SOME EFFECTS OF EXTRA- AND PARALINGUISTIC VARIATION ON THE PHONETIC QUALITY OF VOWELS

Hartmut Traunmüller
Institutionen för lingvistik, Stockholms universitet

ABSTRACT

F_1 at the /i/–/e/ boundary was investigated for synthetic phonated and whispered vowels in which the higher formants and F_0 had been varied. The vowels were presented with and without a leading phrase whose overall F_1 was also varied. The results are discussed in terms of how listeners 'tune in' to a speech signal in order to recover the linguistic information. The effect of cues to this tuning was observed to vary due to interactions as well as spectral and temporal restrictions.

INTRODUCTION

Within the frame of the modulation theory of speech [1], speech perception is seen as a process in which listeners tune in to the 'carrier' (and clock rate) of a speech signal. In order to recover the phonetic quality of vowels they evaluate the deviations of the properties of the signal from those which they expect of a neutral vowel produced by the same speaker with the same vocal effort, in the same voice register, and with the same type of phonation.

In qualitative terms, this theory can explain a large number of experimental results, including the observed dependence of perceived vowel quality on various intrinsic and extrinsic variables, such as F_0 and the formants above F_2 within a vowel or its context. The theory allows for listeners to exploit any kind of cues for tuning in. The experiments to be reported are intended to show which cues actually govern listeners' expectations concerning the frequency position of F_1 when the speaker is unknown and unseen.

For this purpose, the i/e boundary value of F_1 was investigated in a set of experiments which allow conclusions about the contributions of the following variables: F_0 in the vowel itself; F_0 in the vowel and in a leading phrase; F_0 in the leading phrase alone (when the vowel is whispered); whispering vs. phonating; F_1 in the leading phrase; formants above F_1 in the vowel itself; formants above F_1 in the vowel as well as in its leading phrase; and the perceived age and sex of the speaker.

METHOD

Subjects

There were 29 listeners, 15 male and 14 female. They were speakers of standard Swedish, all but one grown up in eastern central Sweden. Nine of them were affiliated to this institute, the rest were university students who were moderately paid.

Stimuli

A female speaker of standard Swedish produced some samples of the phrase *och nu hörs nog snarast V* [ɔ'nʉːhœʂɳʊgsnɒːrast 'Vː] 'and now it's rather likely to hear V' where V stands for the names of the vowels *i* and *e*, pronounced [ij^{ə}] and [e^{ɐ}]. There was always a pause of about 300 ms before the vowel. The utterances were recorded in an anechoic chamber, digitized at 16 kHz, 16 bit/sample, and subjected to LPC analysis with 17 reflection coefficients, 20 ms Hamming window, 5 ms progression. One of the leading phrases and a smoothed interpolation between [ij^{ə}] and [e^{ɐ}] were chosen as models for resynthesis with various modifications of the frequency positions of F_0, F_1, and the formants above F_1, referred to as F_h. The modifications consisted in moving F_0 and/or the F_h into positions typical of male speakers or children. For F_1 of the leading phrase, a normal and a 60 Hz (average) higher or lower position was used. Whispered vowels were synthesized using excitation by white noise, high pass filtered 2nd order Butterworth. In order to keep the number of different stimuli sufficiently small, only 3 or 4 different values of F_1 were used in each condition. In whispering, the i/e boundary was expected at a higher F_1, which is reflected in the choice of F_1 range. The combinations of modifications used are listed in Table 1. The stimuli were recorded on tape using two randomized orders, with pauses of 2.5 s between stimuli, and an additional pause after each sixth stimulus.

Procedure

Three experiments were prepared. In exp. 1, each stimulus was preceded by a phrase whose F_0 and upper formants had been subjected to the same modifications as the test vowel, while its F_1 appeared in normal and in shifted position. Phonated stimuli occurred twice, whispered stimuli once. In exp. 2, the stimuli were presented without a leading phrase, once each. In exp. 3, the different versions of the leading phrase used in exp. 1 were presented for the purpose of classification as to age (child, adolescent, adult, aged) and sex. The three experiments were run in succession, with four to seven subjects at a time.

The stimuli were presented through headphones. The subjects had to identify the vowels by marking the preprinted orthographic symbols *i* /i/, *e* /e/, *ä* /æ/, *y* /y/, *ö* /ø/ or *x* (for any other vowel they heard) on answer sheets. The wording of the leading phrase had been chosen so that all of its vowels were of type *x*, yet with the whole range of variation in F_1 represented. At the beginning of the experimental session, one example stimulus of each of the three experiments was presented for accommodation, without feedback.

RESULTS AND DISCUSSION

Figure 1 shows the identification results obtained in exp. 1 for phonated vowels with unchanged F_0 and F_h presented with an unaltered resynthesis of the original leading phrase. Since we are mainly interested in the distinction between degrees of openness, which can be assumed to be highly correlated with F_1, we are going to neglect the distinction between rounded and spread vowels. Most of the subjects did give some rounded vowel responses, but in this matter, the between-subject agreement was very low, while it was quite high for distinctions in openness or height, i.e., between /i/ and /y/ as opposed to /e/ and /ø/.

It was intended to study the F_1-values at the i/e boundary. Some of these boundaries came, however, to be located outside the range of F_1-variation used. In order to avoid the risk of substantial extrapolation errors, the values presented in the following correspond to the points where 40% of the subjects heard /i/ or /y/ and 60% /e/ and /ø/, with the few other responses neglected. This value will be referred to as 'the boundary value'.

Table 1. Combinations of leding phrases (1st column) and test vowels used. 1st digit = 0: Whispered vowel.

111	111	112	113		012	013	014	
112	111	112	113		012	013	014	
211	211	212	213		012	013	014	
212	211	222	213		012	013	014	
121	121	122	123		022	023	024	
122	121	122	123		022	023	024	
221	221	222	223		022	023	024	
222	221	222	223	224	022	023	024	025
223		222	223	224		023	024	025
322		322	323	324		023	024	025
323		322	323	324		023	024	025
232		232	233	234		033	034	035
233		232	233	234		033	034	035
332		332	333	334		033	034	035
333		332	333	334		033	034	035

Digit value	1	2	3	4	5
1st digit, F0*	0.59	1.00	1.41		
2nd digit, Fh*	0.85	1.00	1.15		
3rd digit, F1	-60	+0	+60	+120	+180

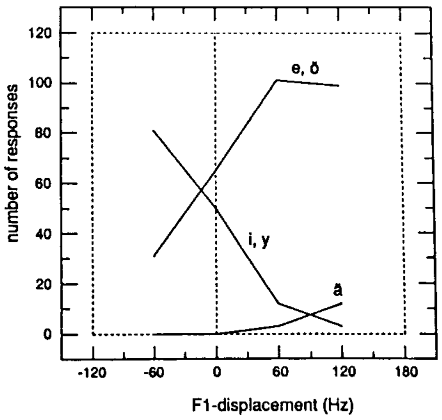


Figure 1. Identifications of phonated vowels with unchanged F_0 and F_h, presented with an unaltered resynthesis of the original leading phrase.

In Figure 2, the boundary value in the whispered stimuli is plotted against that of the otherwise identical phonated stimuli. Evidently, the boundary is situated at higher values of F_1 in the whispered stimuli. This is what could be expected, since F_1 in whispered vowels is usually higher than in the same vowels if phonated. Figure 2 also shows that this difference (in Hz) increases with F_1.

The effects of changes in F_0, F_h, and F_1 of the leading phrase are shown in Figures 3 to 5 for the cases in which the other variables remained unchanged.

The effects of F_0 and F_h were strongly nonlinear. The effect of decreasing F_0 from the original female value was, on average, considerably smaller than that of increasing it, but this can only be asserted for the phonated vowels, or for those cases in which the boundary was no higher than 2.3 units with unchanged F_0. The frequency range within which listeners expect F_1 appears to be limited so that they do not expect it to be shifted further down even if F_0 is lower. For the whispered vowels, this lower limit was not reached.

In contrast, the boundary shift effected by increasing F_h from the original female value was, in absolute terms, about 0.7 units smaller than that of decreasing it. F_h had even a negative effect for the phonated vowels. This can be understood if it is assumed that the variation in the higher formants which listeners are able to utilize for their tuning is restricted to a narrow range which, for the present subjects, did not include what can be found in children. It is fully plausible that this results in a negative effect when the limits of that range are exceeded. It may be that a different result would have been obtained with listeners who are more frequently exposed to speech of children.

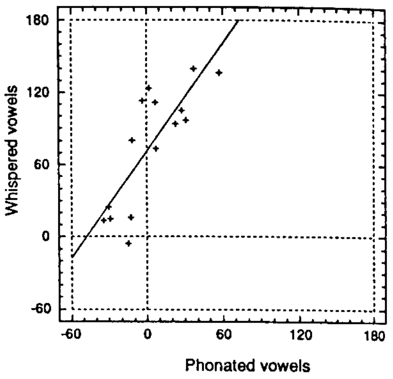
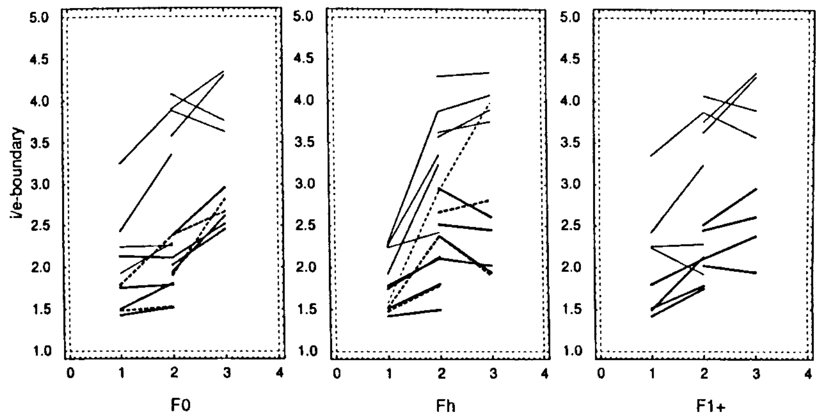


Figure 2. Boundary value (40% i, y; 60% e, ö) of F_1 in the whispered stimuli plotted against that of the otherwise identical phonated stimuli, both presented with leading phrase. Scale values indicate displacement of F_1 from the original value. Regression line also fitted (r = 0.81).

It does not appear to make any difference whether the information on F_0 or F_h is contained only in the vowel itself or both in the vowel and in its leading phrase. There was no significant difference between these cases. The information contained in the vowels themselves was apparently enough to allow the subjects to adapt to the speaker, so that the addition of the leading phrase could not bring about any seizable improvement. This interpretation is supported by the observation that there was a significant and substantial effect of F_0 in the leading phrase when there was no F_0 information in the vowels themselves, i.e., when they were whispered. The latter result also shows that listeners are capable of 'translating' their expectations from voiced to whispered speech, which involves an upward displacement of the i/e boundary.

The effect of F_1 in the leading phrase was, on average, not very large (slope 0.3). It was positive in 12 pairs, while it was negative in one phonated and in three whispered pairs. For voiced vowels, the effect of decreasing F_1 was, nevertheless, larger than that of decreasing F_0. In contrast with the effects of F_0 and F_h, the effects of increases in F_1 were of the same magnitude as those of decreases. The results indicate that F_1 in the leading phrase did not have a very persistent effect, otherwise the slopes in Figure 5 would have been steeper. The pause between the leading phrase and the vowel, and the long duration of the latter have apparently led the subjects to base their expectations more on the intrinsic F_0 and F_h in the vowel itself. The conclusion that listeners attach a lower weight to F_1 of the leading phrase than to F_0 and the formants above F_1 is not likely to hold in general. If the test vowel had been embedded within a phrase without pauses, we would probably have observed a considerably larger effect. This is supported by the informal observation that the phonetic quality of the vowels of the leading phrase itself was not noticeably affected by the frequency displacements in its F_1.



Figures 3 to 5. The effects on the boundary value (40% i, y; 60% e, ö) of F_1 of changes in F_0, F_h (all formants above F_1), and in F_1 of the leading phrase. Cases for which there was no other change in frequency positions connected by lines.

Line types: Thick: Phonated; Thin: Whispered vowels. Full: With a (phonated) leading phrase; Dashed: Isolated vowels.

Scale units as in Table 1.

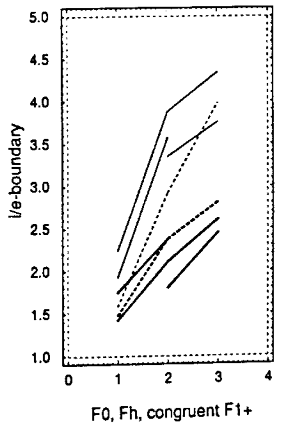


Figure 6 (to the right). Boundaries for the stimuli which simulated modal voice register in a man, a woman, and a child, with original and modified F_1 in the leading phrase. Lines drawn between pairs with congruent values of F_0, F_h, and F_1.

Figure 6 shows the shift in the boundary value for the stimuli which simulated modal voice register in a child, a woman and a man. This figure shows that, for whispered stimuli, the listeners appear to have tuned in themselves fully to the kind of speaker in question, since the slope of the lines is approximately 1.0. With the phonated stimuli, full adaptation to the speaker was apparently not reached, since the slope is smaller, about 0.6. This may have to do with the deficiency in naturalness characteristic of buzz-excited LPC speech.

It remains yet to analyze the possible effects of the listeners' perception of the speaker's age and sex.

CONCLUSIONS

The experiments have shown that all the variables analyzed so far did affect listeners' expectations concerning F_1 at the i/e-boundary, but also that the efficiency of variables is subject to various restrictions in the domains of frequency and time. They have also shown that the amount of a shift in a phoneme boundary brought about by some variable is not a valid measure of its general importance. If the listener is already tuned in to the speaker, there will be no shift at all, no matter how much additional potentially useful information is added.

REFERENCES

[1] H. Traunmüller (1994) *Phonetica 51*, 170–183.